Strict Half-Singleton Bound, Strict Direct Upper Bound for Linear Insertion-Deletion Codes and Optimal Codes

Qinqin Ji, Dabin Zheng, Hao Chen and Xiaoqiang Wang *

Abstract

Insertion-deletion codes (insdel codes for short) are used for correcting synchronization errors in communications, and in other many interesting fields such as DNA storage, date analysis, race-track memory error correction and language processing, and have recently gained a lot of attention. To determine the insdel distances of linear codes is a very challenging problem. The half-Singleton bound on the insdel distances of linear codes due to Cheng-Guruswami-Haeupler-Li is a basic upper bound on the insertion-deletion error-correcting capabilities of linear codes. On the other hand the natural direct upper bound $d_I(\mathcal{C}) \leq 2d_H(\mathcal{C})$ is valid for any insdel code. In this paper, for a linear insdel code $\mathcal{C}$ we propose a strict half-Singleton upper bound $d_I(\mathcal{C}) \leq 2(n-2k+1)$ if $\mathcal{C}$ does not contain the codeword with all 1s, and a stronger direct upper bound $d_I(\mathcal{C}) \leq 2(d_H(\mathcal{C})-t)$ under a weak condition, where $t \geq 1$ is a positive integer determined by the generator matrix. We also give optimal linear insdel codes attaining our strict half-Singleton bound and direct upper bound, and show that the code length of optimal binary linear insdel codes with respect to the (strict) half-Singleton bound is about twice the dimension. Interestingly explicit optimal linear insdel codes attaining the (strict) half-Singleton bound, with the code length being independent of the finite field size, are given.

Keywords: Linear insdel code; strict half-Singleton bound, strict direct upper bound, optimal linear insdel code

1 Introduction

In most communication and storage channels, the most common type of errors are substitution errors, in which a transmitted symbol is replaced with another symbol. However, channels may also suffer from synchronization errors due to slips in synchronization causing the deletion of a symbol from a message or the insertion of an extra symbol into a message [12, 24]. Insdel codes were introduced in [18] for correcting synchronization errors. Insdel errors model also has been widely applied in many interesting fields such as DNA storage, date analysis, race-track memory error correction

and language processing, we refer to [2, 3, 6, 8, 11, 13–15, 17, 19, 21–23, 27] for the construction and application of insdel codes.

For a vector $\mathbf{a} \in \mathbb{F}_q^n$, the support of $\mathbf{a}$ is $\mathrm{supp}(\mathbf{a}) = \{i : a_i \neq 0\}$. The Hamming weight $w_H(\mathbf{a})$ of $\mathbf{a}$ is the number of coordinate positions in its support. The Hamming distance $d_H(\mathbf{a}, \mathbf{b})$ between two vectors $\mathbf{a}$ and $\mathbf{b}$ is defined to be the Hamming weight $w_H(\mathbf{a} - \mathbf{b})$. For a linear code $\mathcal{C} \subset \mathbb{F}_q^n$ of dimension k, its Hamming distance (or weight) d_H is the minimum of Hamming distances $d_H(\mathbf{a}, \mathbf{b})$ between any two different codewords $\mathbf{a}$ and $\mathbf{b}$ in $\mathcal{C}$. It is well-known that the Hamming distance (or weight) of a linear code $\mathcal{C}$ is the minimum Hamming weight of non-zero codewords. The famous Singleton bound $d_H \leq n - k + 1$ is the basic upper bound for linear error-correcting codes. For two codewords $\mathbf{a} \neq \mathbf{b}$ in a code $\mathcal{C} \subset \mathbb{F}_q^n$, the insdel distance $d_I(\mathbf{a}, \mathbf{b})$ between them is defined as the smallest number of insertions and deletions needed to transform one codeword into the other. Similarly to the minimum Hamming distance, the minimum insdel distance of a code is defined as the minimum insdel distance among all its distinct codewords. It is easy to verify that the insdel distance is a metric, and a code or a linear code is called a insdel code or a linear insdel code if we consider insdel metric. The minimal insdel distance of an insdel code is an important parameter, which determines its insertion-deletion error-correcting capability.

The study of insdel codes dates back to the pioneering work of Levenshtein [18]. From then on, the problem to correct the synchronization errors has attracted lots of continuous efforts. For the recent progress in insdel codes, the reader can refer to the nice survey [16] and references therein. Since linear codes have a compact representation, and are efficiently encodable (decodable), we recall the main research progress in linear insdel codes below. In 2010, Abdel-Ghaffar et al. [1] showed that an $[n, k]$ linear code $\mathcal{C}$ over $\mathbb{F}_q$ with $n < 2k$ had the minimum insdel distance $d_I(\mathcal{C}) = 2$, and gave a sufficient and necessary condition for $d_I(\mathcal{C}) = 2$. Actually it was shown in [13] that

$$d_I(\mathbf{a}, \mathbf{b}) = 2(n - \ell),$$

where ℓ is the length of a longest common subsequence of $\mathbf{a}$ and $\mathbf{b}$. It is clear $d_I(\mathbf{a}, \mathbf{b}) \leq 2d_H(\mathbf{a}, \mathbf{b})$ since $\ell \geq n - d_H(\mathbf{a}, \mathbf{b})$ is valid for arbitrary two different vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{F}_q^n$. We call the natural upper bound $d_I(\mathcal{C}) \leq 2d_H(\mathcal{C})$ the direct upper bound for insdel codes. It is true for any insdel code, not only linear insdel codes. Hence it was shown in Haeupler et al. [13] that the minimum insdel distance $d_I(\mathcal{C}) \leq 2(n - k + 1)$ for any $[n, k]$ linear code $\mathcal{C}$ over $\mathbb{F}_q$ from the Singleton bound on the Hamming distances, which is called the direct Singleton bound for linear insdel codes. For insertion-deletion codes the ordering of coordinate positions strongly affects the insdel distances. In this paper we give some upper bounds for insdel distances of linear codes which are valid for any fixed ordering of coordinate positions. There have been many constructions of insdel codes in previous works [1, 4, 7, 9–11, 20, 23, 25, 26].

In [10], Do Duc et al. showed that the minimum insdel distance of any $[n, k]$ Reed-Solomon (RS) code over $\mathbb{F}_q$ is no more than $2n - 2k$ if q is large enough compared to the code length n. Then, Chen et al. [4] generalized this result and showed that for any $[n, k]$ linear code over $\mathbb{F}_q$ with $n > k \geq 2$, the minimum insdel distance is at most $2n - 2k$, and an infinite family of optimal two-dimensional RS codes meeting the bound was constructed. Very recently, Cheng, Gruswami, Haeupler and Li [7] proved the existence of binary linear codes of length n and rate just below $\frac{1}{2}$ capable of correcting $\Omega(n)$ insertions and deletions, and proposed the asymptotic half-Singleton bound for the insdel distances of an $[n, k]$ linear code over $\mathbb{F}_q$, then their results were improved

significantly in [9]. Their half-Singleton bound for an $[n, k]$ linear code $\mathcal{C}$ over $\mathbb{F}_q$ can be reformulated as $d_I(\mathcal{C}) \leq \max\{2(n - 2k + 2), 2\}$. For a simpler proof we refer to [5]. A new coordinate-ordering free upper bound was also given in [5]. It is well-known that the half-Singleton bound is only true for linear insdel codes.

When the dimension $k = 2$, RS codes attaining the bound $2n - 4$ were constructed in [4,7,9]. When the dimension $k \geq 3$, the half-Singleton bound is tighter than $2n - 2k$. Con et al. [9] proved that there were $[n, k]$ RS codes achieving the half-Singleton bound if the field size was large enough and gave a deterministic construction of such codes over much larger fields (of size $n^{k^{O(k)}}$). The code length is small when compared to the field size. As far as we known, up to now there is no explicit construction of optimal $[n, k]$ linear insdel codes attaining the half-Singleton bound for $k \geq 3$.

The direct upper bound $d_I(\mathcal{C}) \leq 2d_H(\mathcal{C})$ is fundamental for insdel codes and the half-Singleton upper bound is fundamental for linear insdel codes. When $d_H \leq n - 2k + 1$, the direct upper bound has to be used to upper bound the insdel distances of codes. In this paper, we show both upper bounds for linear codes can be improved under a weak condition. We first propose a strict half-Singleton upper bound

$$d_I(\mathcal{C}) \leq 2(n - 2k + 1)$$

on the insdel distance for linear insdel codes without the codeword with all 1s by investigating the linear equations associated with the generator matrices. Then, we provide a sufficient condition for a linear insdel code attaining the strict half-Singleton bound, and by this sufficient condition some optimal linear insdel codes with dimension $k \geq 3$ are constructed. Finally, we study the optimal binary linear insdel codes with respect to the (strict) half-Singleton bound and prove that the code length of optimal binary linear insdel codes is about twice the dimension, and conjecture that optimal binary linear insdel codes have parameters $[2k, k, 4]$ or $[2k + 1, k, 4]$ with respect to the half-Singleton bound or the strict half-Singleton bound, respectively. Interestingly explicit optimal linear insdel codes attaining the (strict) half-Singleton bound, with the code length being independent of the finite field size, are obtained. On the other hand we prove that the direct upper bound $d_I(\mathcal{C}) \leq 2d_H(\mathcal{C})$ for arbitrary insdel codes can be improved to $d_I(\mathcal{C}) \leq 2(d_H(\mathcal{C}) - t)$ for linear insdel codes, where $t \geq 1$ is a positive integer determined by the generator matrix. Some examples attaining our strict direct upper bound are given.

The rest of this paper is organized as follows. In section 2, we introduce some definitions of insdel codes and preliminary results on linear insdel codes. Section 3 proposes the strict half-Singleton bound for linear insdel codes and gives another proof of the known half-Singleton bound. In section 4, we give a sufficient condition for constructing optimal linear insdel codes with respect to our strict half-Singleton bound and provide some examples of optimal linear insdel codes. In section 5, we study optimal binary linear insdel codes and show that the code length of optimal binary linear insdel codes is about twice the dimension. In Section 6, we prove the strict direct upper bound and discuss the optimal linear insdel codes attaining this bound. Finally, Section 7 concludes this paper.

2 Preliminaries

Let $\mathbb{F}_q$ be a finite field of q elements and $\mathbb{F}_q^n$ be a vector space over $\mathbb{F}_q$ with dimension n. A subspace $\mathcal{C}$ of $\mathbb{F}_q^n$ over $\mathbb{F}_q$ is called a linear code of length n over $\mathbb{F}_q$. Its dual $\mathcal{C}^\perp$ is a linear code

$\mathcal{C}^\perp = \{(x_1, x_2 \cdots, x_n) \in \mathbb{F}_q^n : \sum_i x_i y_i = 0, \forall (y_1, \ldots, y_n) \in \mathcal{C}\}$. As mentioned above, the Hamming distance of a linear code equals the minimum Hamming weight of its non-zero codewords. A linear code $\mathcal{C}$ is called projective if $d_H(\mathcal{C}^\perp) \geq 3$. That is, any two columns of the generator matrix of $\mathcal{C}$ are linear independent over $\mathbb{F}_q$. The following result shows the property of columns of the generator matrix of a linear code.

Lemma 2.1 *Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$ and denote $d_H(\mathcal{C})$ the minimal Hamming distance of $\mathcal{C}$. Let G denote the generator matrix of $\mathcal{C}$. Then there are k of any $n - d_H(\mathcal{C}) + 1$ columns of G are linearly independent over $\mathbb{F}_q$.*

Proof. Let $s = n - d_H(\mathcal{C}) + 1$ and $G = (G_1, G_2, \cdots, G_n)$, where $G_i = (g_{1i}, g_{2i}, \cdots, g_{ki})^T$. Assume that there exist s columns $G_{j_1}, G_{j_2}, \cdots, G_{j_s}$ of G such that any k vectors of them are linearly dependent over $\mathbb{F}_q$, that is the rank of the matrix $\bar{G} = (G_{j_1}, G_{j_2}, \cdots, G_{j_s})$ is less than k. So, the linear system

$$(x_1, x_2, \cdots, x_k)\bar{G} = \mathbf{0}, \tag{1}$$

has nonzero solutions. Let $\mathbf{y} = (y_1, y_2, \cdots, y_k) \in \mathbb{F}_q^k$ be a nonzero solution of (1) and $\mathbf{c} = \mathbf{y}G$. Then $w_H(\mathbf{c}) \leq n - s = d_H(\mathcal{C}) - 1$. This is a contradiction. $\qquad\square$

In this paper, we mainly consider the insdel distance of linear codes used in high insertions and deletions noise regime. We give the definition of the insdel distance of two vectors as follows.

Definition 2.2 *For two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{F}_q^n$, the insdel distance $d_I(\mathbf{a}, \mathbf{b})$ between $\mathbf{a}$ and $\mathbf{b}$ is the minimal number of insertions and deletions which are needed to transform $\mathbf{a}$ into $\mathbf{b}$. It can be verified that $d_I(\mathbf{a}, \mathbf{b})$ is indeed a metric on $\mathbb{F}_q^n$.*

Let $\mathbf{a} = (a_1, a_2, \cdots, a_n), \mathbf{b} = (b_1, b_2, \cdots, b_n) \in \mathbb{F}_q^n$ be two sequences (or vectors). A common subsequence of $\mathbf{a}$ and $\mathbf{b}$ is a sequence $(c_1, c_2, \cdots, c_m)$ such that $c_s = a_{i_s} = b_{j_s}$ for $1 \leq s \leq m$, $1 \leq i_1 < i_2 < \cdots < i_m \leq n$ and $1 \leq j_1 < j_2 < \cdots < j_m \leq n$. It has been proved that the insdel distance between any two vectors can be characterized by their longest common subsequences.

Lemma 2.3 *[10, Lemma 1] Let $\mathbf{a}, \mathbf{b} \in \mathbb{F}_q^n$. Then we have*

$$d_I(\mathbf{a}, \mathbf{b}) = 2n - 2\ell(\mathbf{a}, \mathbf{b}),$$

where $\ell(\mathbf{a}, \mathbf{b})$ denotes the length of a longest common subsequence of $\mathbf{a}$ and $\mathbf{b}$.

For any two distinct codewords $\mathbf{a}, \mathbf{b} \in \mathcal{C}$, by Lemma 2.3 we know that $d_I(\mathbf{a}, \mathbf{b})$ is even and $d_I(\mathbf{a}, \mathbf{b}) \geq 2$. Like the Hamming distance, the insdel distance of a linear insdel code $\mathcal{C}$ over $\mathbb{F}_q$ is defined as

$$d_I(\mathcal{C}) = \min_{\mathbf{a}, \mathbf{b} \in \mathcal{C}, \mathbf{a} \neq \mathbf{b}} \{ d_I(\mathbf{a}, \mathbf{b}) \}.$$

A linear code $\mathcal{C}$ over $\mathbb{F}_q$ of length n, dimension k and minimum insdel distance $d_I(\mathcal{C})$ is called an $[n, k, d_I(\mathcal{C})]$ linear insdel code over $\mathbb{F}_q$. Like the Hamming metric, an $[n, k, d_I]$ linear insdel code $\mathcal{C}$ has insdel error-correcting capability up to $\lfloor \frac{d_I - 1}{2} \rfloor$ [10]. So, an $[n, k, d_I]$ linear code $\mathcal{C}$ can correct insdel errors if and only if $d_I > 2$. As mentioned in the first section, Chen et al. generalized the Singleton type bound of linear insdel codes [13] to the following case.

4

Lemma 2.4 *[4, Theorem A] Let $\mathcal{C}$ be an $[n,k]$ linear code over $\mathbb{F}_q$ with $n > k > 2$. Then the minimum insdel distance of $\mathcal{C}$ is at most $2n - 2k$, i.e., $d_I(\mathcal{C}) \leq 2n - 2k$.*

Cheng, Gruswami, Haeupler and Li proposed the half-Singleton bound for linear insdel codes in [7]. The non-asymptotic version of half-Singleton bound and a simple proof was given in [5].

Lemma 2.5 (Half-Singleton bound [7]) *Let $\mathcal{C}$ be a non-degenerate linear $[n,k]$ code over $\mathbb{F}_q$. Its insdel distance satisfies*
$$d_I(\mathcal{C}) \leq \max\left\{2(n - 2k + 2),\, 2\right\}.$$

The following lemma shows that a linear insdel code must contain two special codewords if its minimal insdel distance is equal to 2.

Lemma 2.6 *[1, Lemma 1] Let $\mathcal{C}$ be an $[n,k]$ linear code over $\mathbb{F}_q$. Then, $d_I(\mathcal{C}) = 2$ if and only if $\mathcal{C}$ contains a codeword $\mathbf{c} = (c_1, c_2, \cdots, c_n)$ such that, for some $1 \leq u \leq v \leq n$ and $\alpha \in \mathbb{F}_q$, $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ defined by*
$$x_i = \begin{cases} 0, & for\ 1 \leq i < u\ or\ v < i \leq n \\ c_{i+1} - c_i, & for\ u \leq i < v \\ \alpha, & for\ i = v \end{cases}$$

is a nonzero codeword.

3   Strict half-Singleton bound

In this section, we show that the half-Singleton bound on the insdel distance of a linear code $\mathcal{C}$ can be improved if $\mathbf{1} \notin \mathcal{C}$. Based on this improved upper bound, we give another proof of the half-Singleton bound on the minimal insdel distance of linear insdel codes and some useful corollaries. To this end, we first introduce some notation. For positive integers n and s with $s \leq n$, we denote $[n] = \{1, 2, \cdots, n\}$ and $[n]^s$ the set of all vectors of length s whose coordinates are from $[n]$. We say a vector $I = (I_1, I_2, \cdots, I_s) \in [n]^s$ is an increasing vector if its coordinates are monotonically increasing, i.e., for any $u < v$ we have $I_u < I_v$, where I_u is the uth coordinate of I.

Theorem 3.1 (Strict half-Singleton bound) *Let $\mathcal{C}$ be an $[n,k]$ linear code over $\mathbb{F}_q$. If $\mathbf{1} = (1, 1, \cdots, 1) \notin \mathcal{C}$, then the insdel distance of $\mathcal{C}$ satisfies*
$$d_I(\mathcal{C}) \leq \max\left\{2(n - 2k + 1),\, 2\right\}.$$

Proof: Let G denote a generator matrix of $\mathcal{C}$ and d denote the minimum Hamming distance of $\mathcal{C}$. Then there exists a codeword $\mathbf{c} \in \mathcal{C}$ satisfying $w_H(\mathbf{c}) = d$. Thus, $\ell(\mathbf{0}, \mathbf{c}) = n - d$.

Next, we discuss the insdel distance of $\mathcal{C}$ in two cases.

(1) $n - d \geq 2k - 1$. Then, $\ell(\mathbf{0}, \mathbf{c}) = n - d$ and
$$d_I(\mathcal{C}) \leq d_I(\mathbf{0}, \mathbf{c}) = 2(n - \ell(\mathbf{0}, \mathbf{c})) \leq 2(n - 2k + 1).$$

(2) $n - d < 2k - 1$. In the following we show that there exist two different codewords $\mathbf{a}, \mathbf{b} \in \mathcal{C}$ such that $\ell(\mathbf{a}, \mathbf{b}) = n - 1$ or $\ell(\mathbf{a}, \mathbf{b}) \geq 2k - 1$. This implies that

$$d_I(\mathcal{C}) \leq d_I(\mathbf{a}, \mathbf{b}) = 2(n - \ell(\mathbf{a}, \mathbf{b})) \leq \max\{2(n - 2k + 1), 2\}.$$

Let $G = (G_1, G_2, \cdots, G_n)$ be a generator matrix of $\mathcal{C}$, where G_i denote the ith column of G. Let $I = (I_1, I_2, \cdots, I_s)$ and $J = (J_1, J_2, \cdots, J_s)$ be two increasing vectors of $[n]^s$. Define an $2k \times s$ matrix as follows:

$$M_{IJ} = \begin{pmatrix} G_{I_1} & G_{I_2} & \cdots & G_{I_s} \\ G_{J_1} & G_{J_2} & \cdots & G_{J_s} \end{pmatrix}. \tag{2}$$

Consider the linear equations

$$(\mathbf{x}, -\mathbf{y}) M_{IJ} = \mathbf{0}, \tag{3}$$

where $\mathbf{x} = (x_1, x_2, \cdots, x_k)$ and $\mathbf{y} = (y_1, y_2, \cdots, y_k)$. If the system (3) has a nonzero solution $(\mathbf{x}, -\mathbf{y}) \in \mathbb{F}_q^{2k}$ with $\mathbf{x} \neq \mathbf{y}$, then there exist two codewords $\mathbf{a} = (f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_n(\mathbf{x}))$ and $\mathbf{b} = (f_1(\mathbf{y}), f_2(\mathbf{y}), \cdots, f_n(\mathbf{y}))$ in $\mathcal{C}$, where $f_i(\mathbf{x}) = \mathbf{x} G_i$, $f_i(\mathbf{y}) = \mathbf{y} G_i$ for $i = 1, 2, \cdots, n$ such that

$$(f_{I_1}(\mathbf{x}), f_{I_2}(\mathbf{x}), \cdots, f_{I_s}(\mathbf{x})) = (f_{J_1}(\mathbf{y}), f_{J_2}(\mathbf{y}), \cdots, f_{J_s}(\mathbf{y})).$$

This implies that $\ell(\mathbf{a}, \mathbf{b}) \geq s$.

Next, we discuss the solutions to the linear system (3).

Case 1: $n \leq 2k$. We show that there exist two distinct codewords $\mathbf{a}, \mathbf{b} \in \mathcal{C}$ such that $\ell(\mathbf{a}, \mathbf{b}) = n - 1$. Let I, J be any two increasing vectors of $[n]^{n-1}$, that is, $s = n - 1$ in the matrix given in (2). The rank of matrix M_{IJ} defined in (2) is less than $2k$. So, the corresponding linear system (3) has nonzero solutions. Moreover, there exist two increasing vectors $I, J \in [n]^{n-1}$ such that the solution $(\mathbf{x}, -\mathbf{y})$ of the corresponding linear system (3) satisfies $\mathbf{x} \neq \mathbf{y}$. In this case, there exist two distinct codewords $\mathbf{a} = \mathbf{x} G$ and $\mathbf{b} = \mathbf{y} G$ satisfying $\ell(\mathbf{a}, \mathbf{b}) = n - 1$ since $\mathbf{a} \neq \mathbf{b}$. Otherwise, we choose $I = (1, 2, \cdots, n - 1)$ and $J = (2, 3, \cdots, n)$, and let $(\mathbf{x}, -\mathbf{y}) \in \mathbb{F}_q^{2k}$ be a nonzero solution of the linear system (3) with $\mathbf{x} = \mathbf{y}$. This gives

$$(f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_{n-1}(\mathbf{x})) = (f_2(\mathbf{y}), f_3(\mathbf{y}), \cdots, f_n(\mathbf{y})) = (f_2(\mathbf{x}), f_3(\mathbf{x}), \cdots, f_n(\mathbf{x})).$$

So, $f_1(\mathbf{x}) = f_2(\mathbf{x}) = \cdots = f_n(\mathbf{x})$, i.e., $(1, 1, \cdots, 1) \in \mathcal{C}$. This is a contradiction.

Case 2: $n > 2k$. We show that there exist two distinct codewords $\mathbf{a}, \mathbf{b} \in \mathcal{C}$ such that $\ell(\mathbf{a}, \mathbf{b}) \geq 2k - 1$. Let I, J be any two increasing vectors of $[n]^{2k-1}$, that is, $s = 2k - 1$ in the matrix given in (2). The rank of M_{IJ} given in (2) is less than $2k$, and so the corresponding linear system (3) has nonzero solutions. Moreover, there exist two increasing vectors $I, J \in [n]^{2k-1}$ such that the corresponding linear system (3) has a nonzero solution $(\mathbf{x}, -\mathbf{y}) \in \mathbb{F}_q^{2k}$ satisfying $\mathbf{x} \neq \mathbf{y}$. In this case, the code $\mathcal{C}$ has two distinct codewords $\mathbf{a} = \mathbf{x} G$ and $\mathbf{b} = \mathbf{y} G$ in $\mathcal{C}$ satisfying $\ell(\mathbf{a}, \mathbf{b}) \geq 2k - 1$. Otherwise, assume that for any two increasing vectors $I, J \in [n]^{2k-1}$, the corresponding linear system (3) has only solutions with the form $(\mathbf{x}, -\mathbf{x}) \in \mathbb{F}_q^{2k}$. Then we will derive a contradiction.

Choose $I = (1, 2, \cdots, 2k - 1)$ and $J = (2, 3, \cdots, 2k)$. If the corresponding linear system (3) has only nonzero solutions of the form $(\mathbf{x}, -\mathbf{x}) \in \mathbb{F}_q^{2k}$, then we have

$$(f_1(\mathbf{x}), f_2(\mathbf{x}), \cdots, f_{n-1}(\mathbf{x})) = (f_2(\mathbf{y}), f_3(\mathbf{y}), \cdots, f_n(\mathbf{y})) = (f_2(\mathbf{x}), f_3(\mathbf{x}), \cdots, f_n(\mathbf{x})).$$

So, $f_1(\mathbf{x}) = f_2(\mathbf{x}) = \cdots = f_{2k}(\mathbf{x})$. This further shows that the code $\mathcal{C}$ has a codeword of the form

$$\mathbf{c}^{(1)} = (\underbrace{1, 1, \cdots, 1}_{2k}, *, *, \cdots, *).$$

Similarly, choosing $I = (2, 3, \cdots, 2k)$ and $J = (3, 4, \cdots, 2k + 1)$, we can derive that the code $\mathcal{C}$ has a codeword of the form

$$\mathbf{c}^{(2)} = (\star, \underbrace{1, 1, \cdots, 1}_{2k}, \star, \star, \cdots, \star).$$

Repeating the above process $n - 2k + 1$ times, we get the $(n - 2k + 1)$th codeword in $\mathcal{C}$ of the form

$$\mathbf{c}^{(n-2k+1)} = (\diamond, \diamond, \cdots, \diamond, \underbrace{1, 1, \cdots, 1}_{2k}).$$

Suppose that $\mathbf{u}, \mathbf{v} \in \mathbb{F}_q^k$ are the message vectors of $\mathbf{c}^{(1)}$ and $\mathbf{c}^{(2)}$, respectively. Then we have

$$(\mathbf{u}G_2, \mathbf{u}G_3, \cdots, \mathbf{u}G_{2k}) = (\mathbf{v}G_2, \mathbf{v}G_3, \cdots, \mathbf{v}G_{2k}) = (1, 1, \cdots, 1),$$

where $G_i's$ are columns of the generator matrix of G. This implies that

$$(\mathbf{u} - \mathbf{v})\underbrace{(G_2, G_3, \cdots, G_{2k})}_{\bar{G}} = (0, 0, \cdots, 0). \tag{4}$$

Since $n - d < 2k - 1$, i.e., $n - d + 1 \leq 2k - 1$, by Lemma 2.1 there exist k columns of $\bar{G}$ are linearly independent over $\mathbb{F}_q$, i.e., the rank of $\bar{G}$ is equal to k. So, the linear system (4) has only zero solution, i.e., $\mathbf{u} = \mathbf{v}$. This leads to $\mathbf{c}^{(1)} = \mathbf{c}^{(2)}$. Repeating above discussion we derive that the code $\mathcal{C}$ has a codeword

$$\mathbf{c}^{(1)} = \mathbf{c}^{(2)} = \cdots = \mathbf{c}^{(n-2k+1)} = (1, 1, \cdots, 1).$$

This is a contradiction. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ $\square$

Theorem 3.1 and Lemma 2.5 show that the insdel distance of a linear code will be affected by whether it contains the codeword $\mathbf{1}$. This fact can be verified by the following simple example. Let $\mathcal{C}$ be an $[n, 1]$ code over $\mathbb{F}_q$, then $\mathcal{C} = \{a\mathbf{c} : a \in \mathbb{F}_q\}$, where $\mathbf{c} \in \mathbb{F}_q^n$. If $\mathbf{1} \in \mathcal{C}$, then $\mathcal{C} = \{a \cdot \mathbf{1} : a \in \mathbb{F}_q\}$. For any two distinct codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$, $\ell(\mathbf{c}_1, \mathbf{c}_2) = 0$, and so, $d_I(\mathcal{C}) = 2n$. In this case, the insdel distance of the code $\mathcal{C}$ reaches the half-Singleton bound given in Lemma 2.5. If $\mathbf{1} \notin \mathcal{C}$, then we can show that there always exist two codewords $\mathbf{c}_1, \mathbf{c}_2 \in \mathcal{C}$ such that $\ell(\mathbf{c}_1, \mathbf{c}_2) \geq 1$, and so, $d_I(\mathcal{C}) \leq 2(n-1)$. In this case, the insdel distance of the code $\mathcal{C}$ may reach the upper bound given in Theorem 3.1, but never reach the upper bound given in Lemma 2.5.

By Theorem 3.1, we give another proof of the half-Singleton bound on the insdel distance of linear codes.

Corollary 3.2 *Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. Its insdel distance satisfies*

$$d_I(\mathcal{C}) \leq \max\{2(n - 2k + 2), 2\}.$$

Proof: If $\mathbf{1} \notin \mathcal{C}$, by Theorem 3.1, we have $d_I(\mathcal{C}) \leq \max\{2(n - 2k + 1), 2\} \leq \max\{2(n - 2k + 2), 2\}$. If $\mathbf{1} \in \mathcal{C}$, we only need to prove that there exist two distinct codewords $\mathbf{a}, \mathbf{b} \in \mathcal{C}$ satisfying $\ell(\mathbf{a}, \mathbf{b}) \geq \min\{2k - 2, n - 1\}$. Since $\mathbf{1} \in \mathcal{C}$, $\mathcal{C}$ has a generator matrix as the following form:

$$G = \begin{pmatrix} \mathbf{1}_{n-1} & 1 \\ G_{n-1} & \mathbf{0}_{n-1}^T \end{pmatrix}, \tag{5}$$

where G_{n-1} is a $(k - 1) \times (n - 1)$ matrix over $\mathbb{F}_q$, $\mathbf{1}_{n-1} = \underbrace{(1, 1, \cdots, 1)}_{n-1}$ and $\mathbf{0}_{n-1} = \underbrace{(0, 0, \cdots, 0)}_{n-1}$. Let $\mathcal{C}_{n-1}$ be the $[n - 1, k - 1]$ linear code generated by G_{n-1}. If $\mathbf{1}_{n-1} \in \mathcal{C}_{n-1}$, then $\mathbf{1}' = (\mathbf{1}_{n-1}, 0) \in \mathcal{C}$ and $\ell(\mathbf{1}', \mathbf{1}) = n - 1 \geq \min\{2k - 2, n - 1\}$. If $\mathbf{1}_{n-1} \notin \mathcal{C}_{n-1}$, by Theorem 3.1, there exist two distinct codewords $\mathbf{a}_{n-1}, \mathbf{b}_{n-1} \in \mathcal{C}_{n-1}$ such that $\ell(\mathbf{a}_{n-1}, \mathbf{b}_{n-1}) \geq \min\{2k - 3, n - 2\}$. It is clear that $\mathbf{a} = (\mathbf{a}_{n-1}, 0) \in \mathcal{C}$, $\mathbf{b} = (\mathbf{b}_{n-1}, 0) \in \mathcal{C}$, and $\ell(\mathbf{a}, \mathbf{b}) \geq \min\{2k - 2, n - 1\}$. □

For a linear code $\mathcal{C}$ over $\mathbb{F}_q$, we know that $d_I(\mathcal{C}) = 2$ if $n < 2k$ by Lemma 2.5. These codes can not correct insdel errors. When $n = 2k$, from Lemma 4 in [1] we know that the following linear insdel code attains the half-Singleton bound.

Corollary 3.3 *For a positive integer k, let $\mathcal{C}$ be an $[2k, k]$ code over $\mathbb{F}_q$ given by*

$$\mathcal{C} = \{(c_1, c_2, \cdots, c_{2k}) : c_i = c_{2k-i+1} \in \mathbb{F}_q, \ i = 1, 2, \cdots, k\}.$$

Then $d_I(\mathcal{C}) = 4$, i.e., $\mathcal{C}$ is optimal with respect to the half-Singleton bound.

Remark 3.4 *The length of the optimal linear insdel codes given in Corollary 3.3 is independent of the size of the finite field.*

The following corollary shows that only in very special cases, a linear code $\mathcal{C}$ and its dual $\mathcal{C}^\perp$ have the insdel error-correcting capability at the same time. This result directly follows from Theorem 3.1 and Lemma 2.5.

Corollary 3.5 *Let $\mathcal{C}$ be an $[n, k]$ code over $\mathbb{F}_q$ and $\mathcal{C}^\perp$ be its dual code. If both $\mathcal{C}$ and $\mathcal{C}^\perp$ have insdel error-correcting capability, then $n = 2k$. In this case, $\mathbf{1} \in \mathcal{C}$, $d_I(\mathcal{C}) = d_I(\mathcal{C}^\perp) = 4$ and $p \mid n$, where p is the characteristic of the field $\mathbb{F}_q$.*

4 Optimal linear insdel codes attaining the strict half-Singleton bound

In this section, we present a sufficient condition for a linear insdel code to be optimal according to the strict half-Singleton bound given in Theorem 3.1. Then we give several examples of optimal linear insdel codes. To this end, we first introduce some useful notation. Let n, k be positive integers with $2k < n$. Let $I, J \in [n]^{2k}$ be increasing vectors with length $2k$. Let $I \cap J$ be a increasing vector made up of the corresponding equal components of I and J, i.e., $I \cap J = (r_1, r_2, \cdots, r_t)$, $t \leq 2k$, where $r_i = I_{e_i} = J_{e_i}$, which is the e_ith component of I and J for $1 \leq i \leq t$.

Theorem 4.1 *Let $\mathcal{C}$ be an $[n, k]$ code over $\mathbb{F}_q$ with generator matrix $G = (G_1, G_2, \cdots, G_n)$, where G_i is the ith column of G and $n > 2k$. If for every two increasing vectors $I, J \in [n]^{2k}$ with $rank(G_{I \cap J}) < k$, where $G_{I \cap J} = (G_{e_1}, G_{e_2}, \cdots, G_{e_t})$ and $I \cap J = (e_1, e_2, \cdots, e_t)$, it holds that $det(M_{IJ}) \neq 0$, where*

$$M_{IJ} = \begin{pmatrix} G_{I_1} & G_{I_2} & \cdots & G_{I_{2k}} \\ G_{J_1} & G_{J_2} & \cdots & G_{J_{2k}} \end{pmatrix},$$

then $d_I(\mathcal{C}) = 2(n - 2k + 1)$, i.e., $\mathcal{C}$ is optimal with respect to the strict half-Singleton bound.

Proof: First, we show that $\mathbf{1} \notin \mathcal{C}$. Otherwise, assume that $\mathbf{1} \in \mathcal{C}$, then it has a generator matrix as the following form:

$$G' = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ g'_{21} & g'_{22} & \cdots & g'_{2n} \\ \vdots & \vdots & & \vdots \\ g'_{k1} & g'_{k2} & \cdots & g'_{kn} \end{pmatrix} = \begin{pmatrix} G'_1 & G'_2 & \cdots & G'_n \end{pmatrix}.$$

Consider the matrix

$$M'_{IJ} = \begin{pmatrix} G'_{I_1} & G'_{I_2} & \cdots & G'_{I_{2k}} \\ G'_{J_1} & G'_{J_2} & \cdots & G'_{J_{2k}} \end{pmatrix}$$

for two increasing vectors $I, J \in [n]^{2k}$ with $rank(G_{I \cap J}) < k$. Since M'_{IJ} has two rows with all 1s, $det(M'_{IJ}) = 0$. On the other hand, there exists a $k \times k$ invertible matrix Q such that $G = QG'$. Let

$$N = \begin{pmatrix} Q & 0_{k \times k} \\ 0_{k \times k} & Q \end{pmatrix}.$$

It is easy to verify that $NM'_{IJ} = M_{IJ}$. Then $det(M_{IJ}) = det(N) det(M'_{IJ}) = 0$. This is a contradiction. It follows that $\mathbf{1} \notin \mathcal{C}$. By Theorem 3.1, $d_I(\mathcal{C}) \leq 2(n - 2k + 1)$.

Second, we show that for any two different codewords $\mathbf{a}, \mathbf{b} \in \mathcal{C}$, $\ell(\mathbf{a}, \mathbf{b}) \leq 2k - 1$. Otherwise, assume that there exist two distinct codewords $\mathbf{a}, \mathbf{b}$ such that $\ell(\mathbf{a}, \mathbf{b}) \geq 2k$, then there exist two increasing vectors $I, J \in [n]^{2k}$ such that $\mathbf{a}_I = \mathbf{b}_J$, i.e., $\mathbf{a}_{I_s} = \mathbf{b}_{J_s}$ for $s = 1, 2, \cdots, 2k$. Let $\mathbf{x}, \mathbf{y} \in \mathbb{F}_q^k$ be the message symbols of the codewords $\mathbf{a}, \mathbf{b}$ respectively, i.e., $\mathbf{a} = \mathbf{x}G$ and $\mathbf{b} = \mathbf{y}G$. From assumption we see that the linear system

$$\mathbf{z}G_{I \cap J} = (\mathbf{a}_{e_1}, \mathbf{a}_{e_2}, \cdots, \mathbf{a}_{e_t})$$

has two distinct solutions $\mathbf{x}$ and $\mathbf{y}$. So, $rank(G_{I \cap J}) < k$. Since $\mathbf{a}_I = \mathbf{b}_J$, we have

$$(\mathbf{x}, -\mathbf{y})M_{IJ} = 0.$$

So, $det(M_{IJ}) = 0$. This contradicts the assumption in the theorem. Thus, it follows that $d_I(\mathcal{C}) \geq 2(n - 2k + 1)$, and then $d_I(\mathcal{C}) = 2(n - 2k + 1)$ by Theorem 3.1. $\square$

Next we use Theorem 4.1 to give some examples of optimal linear insdel codes.

Example 4.2 *Let $q = 49$ and w be a generator of $\mathbb{F}_q$. Let $\mathcal{C}$ be an $[5, 2]$ code over $\mathbb{F}_q$ with generator matrix*

$$G = \begin{pmatrix} w^{28} & w & w^{39} & w^{26} & w^{20} \\ w^{10} & w^{13} & 2 & w^{37} & w \end{pmatrix}.$$

It is easy to see that any two columns of G are linear independent over $\mathbb{F}_q$. Two different increasing vectors $I, J \in [5]^4$ satisfying $\text{rank}(G_{I \cap J}) < 2$ if and only if $I \cap J = (\emptyset), (1)$ or (5). So, all possible cases of the vectors I and J are as follows: $I = (1,2,3,4)$ and $J = (2,3,4,5)$; $I = (1,2,3,4)$ and $J = (1,3,4,5)$; $I = (1,2,3,5)$ and $J = (2,3,4,5)$. The corresponding matrices M_{IJ} are as follows:

$$M_{IJ} = \begin{pmatrix} w^{28} & w & w^{39} & w^{26} \\ w^{10} & w^{13} & 2 & w^{37} \\ w & w^{39} & w^{26} & w^{20} \\ w^{13} & 2 & w^{37} & w \end{pmatrix}, \begin{pmatrix} w^{28} & w & w^{39} & w^{26} \\ w^{10} & w^{13} & 2 & w^{37} \\ w^{28} & w^{39} & w^{26} & w^{20} \\ w^{10} & 2 & w^{37} & w \end{pmatrix}, \begin{pmatrix} w^{28} & w & w^{39} & w^{20} \\ w^{10} & w^{13} & 2 & w \\ w & w^{39} & w^{26} & w^{20} \\ w^{13} & 2 & w^{37} & w \end{pmatrix}.$$

By help of Magma one easily show that $\det(M_{IJ}) \neq 0$ for above three cases. From Theorem 4.1 we know that $d_I(\mathcal{C}) = 4$. In fact, we find two codewords $\mathbf{a} = (w^{38}, w^{14}, w^7, w^{15}, w^{21})$, $\mathbf{b} = (w^2, w^{38}, w^{14}, w^7, w^2)$ satisfying $\ell(\mathbf{a}, \mathbf{b}) = 3$, and so $d_I(\mathbf{a}, \mathbf{b}) = 4$.

Example 4.3 Let $q = 121$, and w be a generator of $\mathbb{F}_q$. Let $\mathcal{C}$ be an $[8,3]$ code over $\mathbb{F}_q$ with generator matrix

$$G = \begin{pmatrix} w^{40} & w^{20} & w^{22} & w^3 & w^{49} & w^{55} & w^{54} & w^{65} \\ w^{86} & w^{27} & w^{89} & w^{64} & w^{73} & w^{23} & w^{44} & w^{79} \\ w^{88} & w^{103} & w^{110} & w^{97} & w^{21} & w^{51} & w^{47} & w^{70} \end{pmatrix}.$$

By help of Magma, we can verify that $\det(M_{IJ}) \neq 0$ for all two different increasing vectors $I, J \in [8]^4$ that satisfy $\text{rank}(G_{I \cap J}) < 3$. From Theorem 4.1, we know that $d_I(\mathcal{C}) = 6$. In fact, we find two codewords $\mathbf{a} = (w^{95}, w, w^2, w^{80}, w^{67}, w^{40}, w^{31}, w^{79})$, $\mathbf{b} = (6, w^{95}, w, w^2, w^{80}, w^{67}, w^6, w^{112})$ satisfying $\ell(\mathbf{a}, \mathbf{b}) = 5$, and so $d_I(\mathbf{a}, \mathbf{b}) = 6$.

Example 4.4 Let $q = 169$, and w be a generator of $\mathbb{F}_q$. Let $\mathcal{C}$ be an $[9,4]$ code over $\mathbb{F}_q$ with generator matrix

$$G = \begin{pmatrix} w^{81} & w^{120} & w^4 & w^{136} & w^{147} & w^{71} & w^{166} & w^{132} & w^{103} \\ w^{83} & w^{155} & w^{82} & w^{163} & w^{48} & w^{36} & w^{88} & w^{63} & w^{45} \\ w^{143} & w^{85} & w^{72} & w^{146} & w^{117} & w^{18} & w^{95} & w^{12} & w^{134} \\ w^{131} & w^{160} & w^{27} & w^{148} & w^{164} & w^7 & w^{109} & w^{107} & w^{32} \end{pmatrix}.$$

By help of Magma, we can verify that $\det(M_{IJ}) \neq 0$ for all two different increasing vectors $I, J \in [9]^4$ that satisfy $\text{rank}(G_{I \cap J}) < 4$. From Theorem 4.1, we know $d_I(\mathcal{C}) = 4$. In fact, we find two codewords $\mathbf{a} = (w^9, w^{127}, w^{13}, w^{22}, w^{21}, w^{11}, w^{53}, w^{165}, w^{110})$, $\mathbf{b} = (w^{120}, w^9, w^{127}, w^{13}, w^{22}, w^{21}, w^{11}, w^{53}, 7)$ satisfying $\ell(\mathbf{a}, \mathbf{b}) = 7$, and so $d_I(\mathbf{a}, \mathbf{b}) = 4$.

In the following, we present a class of optimal $[2k+1, k]$ linear insdel codes over $\mathbb{F}_q$ for some positive integer k. To this end, we first introduce some notation. Let t be a positive integer with $t \leq k$ and $\Omega_t = \{t, t+1, \cdots, k\}$. Denote by $\Omega_t^o = \{i \in \Omega_t \mid k - i \text{ is odd }\}$, $\Omega_t^e = \{i \in \Omega_t \mid k - i \text{ is even }\}$ and

$$G = \begin{pmatrix} 1 & 0 & \cdots & 0 & a_1 & 0 & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 0 & a_2 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & a_k & 1 & \cdots & 0 & 0 \end{pmatrix}_{k \times (2k+1)}, \tag{6}$$

where $a_i \in \mathbb{F}_q$ for $i = 1, 2, \cdots, k$ satisfy $\sum_{i=1}^k a_i \neq 1$.

Proposition 4.5 *Let symbols be given as above and $\mathcal{C}$ be an $[2k+1, k]$ code with generator matrix G given in (6). If for any t with $1 \le t \le k$ satisfies*

$$\sum_{i \in \Omega_t^o} a_i - \sum_{i \in \Omega_t^e} a_i \neq 1,$$

then $d_I(\mathcal{C}) = 4$, i.e., $\mathcal{C}$ is optimal with respect to the strict half-Singleton bound.

Proof: Since $\sum_{i=1}^k a_i \neq 1$, it is easy to verify that $\mathbf{1} \notin \mathcal{C}$. So, $d_I(\mathcal{C}) \le 2(2k+1-2k+1) = 4$ by Theorem 3.1. In the following, we show that $d_I(\mathcal{C}) \neq 2$.

Assume $d_I(\mathcal{C}) = 2$, then the linear code $\mathcal{C}$ contains a codeword $\mathbf{c} = (c_1, c_2, \cdots, c_n)$ and a nonzero codeword $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ as characterized in Lemma 2.6 for some u, v and α, where $n = 2k+1$. Since $\mathbf{x}$ is nonzero, from the representation of codewords in $\mathcal{C}$ we know that $x_i \neq 0$ for some $i < k+1$. Thus $u < k+1$ and $v > k+1$. Let $t = \min\{i \mid i \in [n] \text{ and } x_i \neq 0\}$ and $t' = \max\{i \mid i \in [n] \text{ and } x_i \neq 0\}$, then $t' = n - t + 1$, $1 \le t \le k$ and $u \le t < k+1 < t' \le v$. Thus,

$$x_i = \begin{cases} 0, & for\ 1 \le i < t \ or\ t' < i \le n \\ c_{i+1} - c_i, & for\ t \le i < t' \\ \beta, & for\ i = t' \end{cases},$$

where $\beta \in \mathbb{F}_q$. From the generator matrix of $\mathcal{C}$ we know that the codeword $\mathbf{c} = (c_1, c_2, \cdots, c_n) \in \mathcal{C}$ satisfies that

$$c_{k+1} = \sum_{i=1}^k a_i c_i, \ c_j = c_{n+1-j}, \ j = k+2, k+3, \cdots, n. \tag{7}$$

Then the codeword $\mathbf{y} = \mathbf{c} + \mathbf{x} = (y_1, y_2, \cdots, y_n) \in \mathcal{C}$ satisfies that

$$y_i = \begin{cases} c_i, & for\ 1 \le i < t \ or\ t' < i \le n \\ c_{i+1}, & for\ t \le i < t' \\ c_t + \beta, & for\ i = t' \end{cases}.$$

Since any codeword in $\mathcal{C}$ satisfies the relation given in (7), from the representation of $\mathbf{y}$ we have that

$$c_t = c_{t+2} = \cdots = c_{k-1} = c_{k+1} \text{ and } c_{t+1} = c_{t+3} = \cdots = c_{k-2} = c_k$$

if $k - t$ is odd, and

$$c_t = c_{t+2} = \cdots = c_{k-2} = c_k \text{ and } c_{t+1} = c_{t+3} = \cdots = c_{k-1} = c_{k+1}$$

if $k - t$ is even. So, for the codeword $\mathbf{c}$ we have

$$c_{k+1} = \sum_{i=1}^{t-1} c_i a_i + c_k \sum_{i \in \Omega_t^e} a_i + c_{k+1} \sum_{i \in \Omega_t^o} a_i, \tag{8}$$

and for the codeword $\mathbf{y}$ we have

$$c_k = \sum_{i=1}^{t-1} c_i a_i + c_k \sum_{i \in \Omega_t^o} a_i + c_{k+1} \sum_{i \in \Omega_t^e} a_i. \tag{9}$$

11

By $(8) - (9)$, we have

$$c_{k+1} - c_k = (c_{k+1} - c_k)\left(\sum_{i\in\Omega_t^o} a_i - \sum_{i\in\Omega_t^e} a_i\right). \tag{10}$$

Since $\mathbf{x}$ is a nonzero codeword, we know that $c_{k+1} \neq c_k$. From (10) we have

$$\sum_{i\in\Omega_t^o} a_i - \sum_{i\in\Omega_t^e} a_i = 1.$$

This is a contradiction. So, $d_I(\mathcal{C}) = 4$, and $\mathcal{C}$ is an optimal linear insdel code with respect to the strict half-Singleton bound.

Remark 4.6 *When $q > 2$, one can verify that for any t with $1 \leq t \leq k$, there exist $a_i \in \mathbb{F}_q$ satisfying*

$$\sum_{i=1}^{k} a_i \neq 1 \text{ and } \sum_{i\in\Omega_t^o} a_i - \sum_{i\in\Omega_t^e} a_i \neq 1. \tag{11}$$

For example, $a_{k-1} \in \mathbb{F}_q \setminus \{0,1\}$ and $a_i = 0$ for all $i \in \{1,2,\cdots,k\} \setminus \{k-1\}$ satisfy (11). So, there exist optimal $[2k+1,k,4]$ linear insdel codes over $\mathbb{F}_q$ if $q > 2$. Moreover, the length of $\mathcal{C}$ is independent of the size of the finite field $\mathbb{F}_q$.

5 Optimal binary linear insdel codes attaining the (strict) half-Singleton bound

In this section we study optimal linear insdel codes over $\mathbb{F}_2$ with respect to the half-Singleton bound and the strict half-Singleton bound proposed in Theorem 3.1, respectively.

Lemma 5.1 *For a positive integer k, let $\mathcal{C}$ be an $[2k+3,k]$ linear insdel code over $\mathbb{F}_2$ without codeword with $2k$ consecutive coordinates being 1. Then there exist two distinct codewords $\mathbf{u}, \mathbf{v} \in \mathcal{C}$ such that $\ell(\mathbf{u},\mathbf{v}) \geq 2k$.*

Proof: Let d_H be the minimal Hamming distance of $\mathcal{C}$, then there exist a codeword $\mathbf{z} \in \mathcal{C}$ such that $w_H(\mathbf{z}) = d_H$. So $\ell(\mathbf{z},\mathbf{0}) = 2k+3-d_H$. If $d_H \leq 3$, then the conclusion follows. Next we discuss the case of $d_H > 3$.

Let $G = (G_1, G_2, \cdots, G_{2k+3})$ be a generator matrix of $\mathcal{C}$. Consider the linear equations

$$(\mathbf{x}, -\mathbf{y})\underbrace{\begin{pmatrix} G_1 & G_2 & \cdots & G_{2k-1} & \\ G_2 & G_3 & \cdots & & G_{2k} \end{pmatrix}}_{M} = 0. \tag{12}$$

The rank of M is less than $2k$. So, the linear system (12) has a nonzero solution $(\mathbf{x_1}, -\mathbf{y_1}) \in \mathbb{F}_2^{2k}$. Moreover, we claim that $\mathbf{x_1} \neq \mathbf{y_1}$. Otherwise, if $\mathbf{x_1} = \mathbf{y_1}$, then from (12) we have

$$(f_1(\mathbf{x_1}), f_2(\mathbf{x_1}), \cdots, f_{2k-1}(\mathbf{x_1})) = (f_2(\mathbf{y_1}), f_3(\mathbf{y_1}), \cdots, f_{2k}(\mathbf{y_1})) = (f_2(\mathbf{x_1}), f_3(\mathbf{x_1}), \cdots, f_{2k}(\mathbf{x_1})),$$

where $f_i(\mathbf{x}) = \mathbf{x}G_i$ for $1 \leq i \leq 2k$. So, $f_1(\mathbf{x_1}) = f_2(\mathbf{x_1}) = \cdots = f_{2k}(\mathbf{x_1})$. Since $d_H(\mathcal{C}) > 3$, we derive that $\mathcal{C}$ has a codeword of the form $(\underbrace{1, 1, \cdots, 1}_{2k}, *, *, *)$. This is a contradiction.

Let $\mathbf{x_1}, \mathbf{y_1}$ be message symbols of codewords $\mathbf{a}, \bar{\mathbf{a}}$, respectively. Then they are different and have the following form:

$$\mathbf{a} = (a_1, a_2, a_3, a_4, \cdots, a_{2k-1}, \alpha_1, \alpha_2, \alpha_3, \alpha_4), \ \bar{\mathbf{a}} = (\bar{\alpha}_1, a_1, a_2, a_3, \cdots, a_{2k-2}, a_{2k-1}, \bar{\alpha}_2, \bar{\alpha}_3, \bar{\alpha}_4).$$

It is clear that the length of the longest common subsequence of $\mathbf{a}$ and $\bar{\mathbf{a}}$ is at least $2k - 1$. If there exist some $i \in \{1, 2, 3, 4\}$ and $j \in \{2, 3, 4\}$ such that $\alpha_i = \bar{\alpha}_j$, then $\ell(\mathbf{a}, \bar{\mathbf{a}}) \geq 2k$, and the conclusion follows. If $\alpha_i \neq \bar{\alpha}_j$ for any $i \in \{1, 2, 3, 4\}$ and $j \in \{2, 3, 4\}$, we have

$$\alpha_i + \bar{\alpha}_j = 1. \tag{13}$$

By choosing proper matrix M as in (12) we can show that $\mathcal{C}$ has two different codewords $\mathbf{b}$ and $\bar{\mathbf{b}}$ as the following form:

$$\mathbf{b} = (\beta_1, b_1, b_2, b_3, \cdots, b_{2k-2}, b_{2k-1}, \beta_2, \beta_3, \beta_4), \ \bar{\mathbf{b}} = (\bar{\beta}_1, \bar{\beta}_2, b_1, b_2, \cdots, b_{2k-3}, b_{2k-2}, b_{2k-1}, \bar{\beta}_3, \bar{\beta}_4).$$

It is clear that the length of the longest common subsequence of $\mathbf{b}$ and $\bar{\mathbf{b}}$ is at least $2k - 1$. If $\beta_1 \in \{\bar{\beta}_1, \bar{\beta}_2\}$ or $\{\beta_2, \beta_3, \beta_4\} \cap \{\bar{\beta}_3, \bar{\beta}_4\} \neq \emptyset$, then $\ell(\mathbf{b}, \bar{\mathbf{b}}) \geq 2k$, and the conclusion follows. Otherwise, for $j \in \{1, 2\}$, $u \in \{2, 3, 4\}$ and $v \in \{3, 4\}$ we have

$$\beta_1 + \bar{\beta}_j = 1 \text{ and } \beta_u + \bar{\beta}_v = 1. \tag{14}$$

Since $\mathcal{C}$ is a linear code, $\mathbf{a} + \mathbf{b}$ and $\bar{\mathbf{a}} + \bar{\mathbf{b}}$ are also codewords of $\mathcal{C}$ and have the following form:

$$\mathbf{a} + \mathbf{b} = (a_1 + \beta_1, a_2 + b_1, a_3 + b_2, a_4 + b_3, \cdots, a_{2k-1} + b_{2k-2}, \alpha_1 + b_{2k-1}, \alpha_2 + \beta_2, \alpha_3 + \beta_3, \alpha_4 + \beta_4),$$

$$\bar{\mathbf{a}} + \bar{\mathbf{b}} = (\bar{\alpha}_1 + \bar{\beta}_1, a_1 + \bar{\beta}_2, a_2 + b_1, a_3 + b_2, \cdots, a_{2k-2} + b_{2k-3}, a_{2k-1} + b_{2k-2}, \bar{\alpha}_2 + b_{2k-1}, \bar{\alpha}_3 + \bar{\beta}_3, \bar{\alpha}_4 + \bar{\beta}_4).$$

If $\mathbf{a} + \mathbf{b} \neq \bar{\mathbf{a}} + \bar{\mathbf{b}}$, by (13) and (14) we have that $\ell(\mathbf{a} + \mathbf{b}, \bar{\mathbf{a}} + \bar{\mathbf{b}}) \geq 2k$, and the conclusion follows. If $\mathbf{a} + \mathbf{b} = \bar{\mathbf{a}} + \bar{\mathbf{b}}$, by (13) and (14), then we can derive that $\mathbf{a} + \mathbf{b} = (0, \underbrace{1, 1, \cdots, 1}_{2k-1}, 0, 0, 0)$ or $(1, \underbrace{0, 0, \cdots, 0}_{2k-1}, 1, 1, 1)$.

Again by similar analysis of the beginning of the proof, we can show that $\mathcal{C}$ has two different codewords $\mathbf{c}$ and $\bar{\mathbf{c}}$ as the following form:

$$\mathbf{c} = (\gamma_1, \gamma_2, c_1, c_2, c_3, \cdots, c_{2k-2}, c_{2k-1}, \gamma_3, \gamma_4), \ \bar{\mathbf{c}} = (\bar{\gamma}_1, \bar{\gamma}_2, \bar{\gamma}_3, c_1, c_2, \cdots, c_{2k-3}, c_{2k-2}, c_{2k-1}, \bar{\gamma}_4).$$

It is clear that the length of the longset common subsequence of $\mathbf{c}$ and $\bar{\mathbf{c}}$ is at least $2k - 1$. If $\{\gamma_1, \gamma_2\} \cap \{\bar{\gamma}_1, \bar{\gamma}_2, \bar{\gamma}_3\} \neq \emptyset$ or $\bar{\gamma}_4 \in \{\gamma_3, \gamma_4\}$, then $\ell(\mathbf{c}, \bar{\mathbf{c}}) \geq 2k$, and the conclusion follows. Otherwise, for $i \in \{3, 4\}$, $s \in \{1, 2\}$ and $t \in \{1, 2, 3\}$ we have

$$\gamma_i + \bar{\gamma}_4 = 1 \text{ and } \gamma_s + \bar{\gamma}_t = 1. \tag{15}$$

Since $\mathcal{C}$ is a linear code, $\mathbf{b} + \mathbf{c}$ and $\bar{\mathbf{b}} + \bar{\mathbf{c}}$ are also codewords of $\mathcal{C}$ and have the following form:

$$\mathbf{b} + \mathbf{c} = (\beta_1 + \gamma_1, b_1 + \gamma_2, b_2 + c_1, b_3 + c_2, \cdots, b_{2k-1} + c_{2k-2}, \beta_2 + c_{2k-1}, \beta_3 + \gamma_3, \beta_4 + \gamma_4)$$

13

and

$$\bar{\mathbf{b}} + \bar{\mathbf{c}} = (\bar{\beta}_1 + \bar{\gamma}_1, \bar{\beta}_2 + \bar{\gamma}_2, b_1 + \bar{\gamma}_3, b_2 + c_1, b_3 + c_2, \cdots, b_{2k-1} + c_{2k-2}, \bar{\beta}_3 + c_{2k-1}, \bar{\beta}_4 + \bar{\gamma}_4).$$

If $\mathbf{b} + \mathbf{c} \neq \bar{\mathbf{b}} + \bar{\mathbf{c}}$, by (14) and (15) we have that $\ell(\mathbf{b} + \mathbf{c}, \bar{\mathbf{b}} + \bar{\mathbf{c}}) \geq 2k$ and the conclusion follows. If $\mathbf{b} + \mathbf{c} = \bar{\mathbf{b}} + \bar{\mathbf{c}}$, by (14) and (15) we can derive that $\mathbf{b} + \mathbf{c} = (0, 0, \underbrace{1, 1, \cdots, 1}_{2k-1}, 0, 0)$ or $(1, 1, \underbrace{0, 0, \cdots, 0}_{2k-1}, 1, 1)$.

If $\mathbf{a} + \mathbf{b} = \bar{\mathbf{a}} + \bar{\mathbf{b}}$ and $\mathbf{b} + \mathbf{c} = \bar{\mathbf{b}} + \bar{\mathbf{c}}$ at the same time, then we can derive that $\ell(\mathbf{a} + \mathbf{b}, \mathbf{b} + \mathbf{c}) = 2k + 2$ or $\ell(\mathbf{a} + \mathbf{c}, \mathbf{b} + \mathbf{c}) = 2k$ or $\ell(\mathbf{a} + \mathbf{b}, \mathbf{a} + \mathbf{c}) = 2k$. So, we can always obtain two distinct codewords of $\mathcal{C}$ such that the length of their longest common subsequence is at least $2k$. $\qquad \square$

Lemma 5.2 *For a positive integer k, let $\mathcal{C}$ be an $[2k + 3, k]$ linear insdel code over $\mathbb{F}_2$ having a codeword with $2k$ consecutive coordinates being 1. Then there exist two distinct codewords $\mathbf{u}, \mathbf{v} \in \mathcal{C}$ such that $\ell(\mathbf{u}, \mathbf{v}) \geq 2k$.*

Proof: Let d_H be the minimal Hamming distance of $\mathcal{C}$, then there exist a codeword $\mathbf{c} \in \mathcal{C}$ such that $w_H(\mathbf{c}) = d_H$. So, $\ell(\mathbf{c}, \mathbf{0}) = 2k + 3 - d_H$. If $d_H \leq 3$, then the conclusion follows. Next, we discuss the case of $d_H > 3$.

We only prove the case that $\mathcal{C}$ has a codeword of the form $\mathbf{h} = (\underbrace{1, 1, \cdots, 1}_{2k}, 0, 0, 0)$, and the other cases can be shown similarly.

Let $G = (G_1, G_2, \cdots, G_{2k+3})$ be a generator matrix of $\mathcal{C}$. Consider the linear equations

$$(\mathbf{x}, -\mathbf{y}) \underbrace{\begin{pmatrix} G_4 & G_5 & \cdots & G_{2k+2} \\ G_5 & G_6 & \cdots & G_{2k+3} \end{pmatrix}}_{M} = 0. \tag{16}$$

The rank of M is less than $2k$. So, the linear system (16) has a nonzero solution $(\mathbf{x_1}, -\mathbf{y_1}) \in \mathbb{F}_2^{2k}$. If $\mathbf{x_1} = \mathbf{y_1}$ then we have

$$(f_4(\mathbf{x_1}), f_5(\mathbf{x_1}), \cdots, f_{2k+2}(\mathbf{x_1})) = (f_5(\mathbf{y_1}), f_6(\mathbf{y_1}), \cdots, f_{2k+3}(\mathbf{y_1})) = (f_5(\mathbf{x_1}), f_6(\mathbf{x_1}), \cdots, f_{2k+3}(\mathbf{x_1})).$$

where $f_i(\mathbf{x}) = \mathbf{x} G_i$ for $4 \leq i \leq 2k + 3$. So, $f_4(\mathbf{x_1}) = f_5(\mathbf{x_1}) = \cdots = f_{2k+3}(\mathbf{x_1})$. Since $d_H > 3$, we derive that $\mathcal{C}$ has a codeword of the form $\mathbf{a} = (*, *, *, \underbrace{1, 1, \cdots, 1}_{2k})$. It is clear that $\mathbf{a} \neq \mathbf{h}$ and $\ell(\mathbf{a}, \mathbf{h}) \geq 2k$, then the conclusion follows. If $\mathbf{x_1} \neq \mathbf{y_1}$, and let $\mathbf{a} = \mathbf{x_1} G$ and $\bar{\mathbf{a}} = \mathbf{y_1} G$. Then $\mathbf{a}$ and $\bar{\mathbf{a}}$ have the following form:

$$\mathbf{a} = (\alpha_1, \alpha_2, \alpha_3, a_1, a_2, a_3, a_4, \cdots, a_{2k-1}, \alpha_4), \quad \bar{\mathbf{a}} = (\bar{\alpha}_1, \bar{\alpha}_2, \bar{\alpha}_3, \bar{\alpha}_4, a_1, a_2, a_3, \cdots, a_{2k-2}, a_{2k-1}).$$

It is obvious that the length of the longest common subsequence of $\mathbf{a}$ and $\bar{\mathbf{a}}$ is at least $2k - 1$. If there exist some $i \in \{1, 2, 3\}$ and $j \in \{1, 2, 3, 4\}$ such that $\alpha_i = \bar{\alpha}_j$, then $\ell(\mathbf{a}, \bar{\mathbf{a}}) \geq 2k$, and the conclusion follows. Otherwise, for any $i \in \{1, 2, 3\}$ and $j \in \{1, 2, 3, 4\}$, we have

$$\alpha_i + \bar{\alpha}_j = 1. \tag{17}$$

14

By choosing proper matrices M as in (16), we can show that $\mathcal{C}$ has four codewords $\mathbf{b}, \bar{\mathbf{b}}, \mathbf{c}$ and $\bar{\mathbf{c}}$ as the following form:

$$\mathbf{b} = (\beta_1, \beta_2, b_1, b_2, b_3, b_4, \cdots, b_{2k-2}, b_{2k-1}, \beta_3, \beta_4), \ \ \bar{\mathbf{b}} = (\bar{\beta}_1, \bar{\beta}_2, \bar{\beta}_3, b_1, b_2, b_3, \cdots, b_{2k-3}, b_{2k-2}, b_{2k-1}, \bar{\beta}_4);$$

$$\mathbf{c} = (\gamma_1, \gamma_2, c_1, c_2, c_3, c_4, \cdots, c_{2k-2}, \gamma_3, c_{2k-1}, \gamma_4), \ \ \bar{\mathbf{c}} = (\bar{\gamma}_1, \bar{\gamma}_2, \bar{\gamma}_3, c_1, c_2, c_3, \cdots, c_{2k-2}, \bar{\gamma}_4, c_{2k-1}).$$

By analysis similar to that above, we have $\ell(\mathbf{b}, \mathbf{h}) \geq 2k$ if $\mathbf{b} = \bar{\mathbf{b}}$. Otherwise, if $\mathbf{b} \neq \bar{\mathbf{b}}$, then the length of the longest common subsequence of $\mathbf{b}$ and $\bar{\mathbf{b}}$ is at least $2k-1$. Only when $\{\beta_1, \beta_2\} \cap \{\bar{\beta}_1, \bar{\beta}_2, \bar{\beta}_3\} = \emptyset$ and $\bar{\beta}_4 \notin \{\beta_3, \beta_4\}$, we have $\ell(\mathbf{b}, \bar{\mathbf{b}}) = 2k - 1$. In this case, for $i \in \{1, 2\}$, $j \in \{1, 2, 3\}$ and $u \in \{3, 4\}$, we have

$$\beta_i + \bar{\beta}_j = 1 \text{ and } \beta_u + \bar{\beta}_4 = 1. \tag{18}$$

Since $\mathcal{C}$ is linear, $\mathbf{a} + \mathbf{b}$ and $\bar{\mathbf{a}} + \bar{\mathbf{b}}$ are codewords of $\mathcal{C}$ and have the following form:

$$\mathbf{a} + \mathbf{b} = (\alpha_1 + \beta_1, \alpha_2 + \beta_2, \alpha_3 + b_1, a_1 + b_2, a_2 + b_3, \cdots, a_{2k-2} + b_{2k-1}, a_{2k-1} + \beta_3, \alpha_4 + \beta_4),$$

$$\bar{\mathbf{a}} + \bar{\mathbf{b}} = (\bar{\alpha}_1 + \bar{\beta}_1, \bar{\alpha}_2 + \bar{\beta}_2, \bar{\alpha}_3 + \bar{\beta}_3, \bar{\alpha}_4 + b_1, a_1 + b_2, \cdots, a_{2k-3} + b_{2k-2}, a_{2k-2} + b_{2k-1}, a_{2k-1} + \bar{\beta}_4).$$

If $\mathbf{a} + \mathbf{b} \neq \bar{\mathbf{a}} + \bar{\mathbf{b}}$, by (17) and (18), we have $\ell(\mathbf{a} + \mathbf{b}, \bar{\mathbf{a}} + \bar{\mathbf{b}}) \geq 2k$ and the conclusion follows. If $\mathbf{a} + \mathbf{b} = \bar{\mathbf{a}} + \bar{\mathbf{b}}$, we can derive that $\mathbf{a} + \mathbf{b} = (0, 0, 0, \underbrace{1, 1, \cdots, 1}_{2k-1}, 0)$ or $(1, 1, 1, \underbrace{0, 0, \cdots, 0}_{2k-1}, 1)$.

Next, we consider the codewords $\mathbf{c}$ and $\bar{\mathbf{c}}$. If $\mathbf{c} = \bar{\mathbf{c}}$, then we derive that $\mathbf{c} = (*, *, \underbrace{1, 1, \cdots, 1}_{2k+1})$ or $(*, *, \underbrace{1, 1, \cdots, 1}_{2k-1}, 0, 0)$ or $(1, 1, \underbrace{0, 0, \cdots, 0}_{2k-1}, 1, 1)$. In this case, we have $\ell(\mathbf{c}, \mathbf{h}) \geq 2k$ or $\ell(\mathbf{a} + \mathbf{b}, \mathbf{c}) \geq 2k$ or $\ell(\mathbf{a} + \mathbf{b} + \mathbf{c}, \mathbf{h}) \geq 2k$, then conclusion follows. If $\mathbf{c} \neq \bar{\mathbf{c}}$, it is clear that the length of the longest common subsequence of $\mathbf{c}$ and $\bar{\mathbf{c}}$ is at least $2k - 1$. Only when $\{\gamma_1, \gamma_2\} \cap \{\bar{\gamma}_1, \bar{\gamma}_2, \bar{\gamma}_3\} = \emptyset$ and $\gamma_3 \neq \bar{\gamma}_4$, i.e., for $i \in \{1, 2\}$ and $j \in \{1, 2, 3\}$,

$$\gamma_i + \bar{\gamma}_j = 1 \text{ and } \gamma_3 + \bar{\gamma}_4 = 1, \tag{19}$$

we have $\ell(\mathbf{c}, \bar{\mathbf{c}}) = 2k - 1$. In this case, we consider the codewords $\mathbf{a} + \mathbf{c}, \bar{\mathbf{a}} + \bar{\mathbf{c}} \in \mathcal{C}$ as the following form:

$$\mathbf{a} + \mathbf{c} = (\alpha_1 + \gamma_1, \alpha_2 + \gamma_2, \alpha_3 + c_1, a_1 + c_2, a_2 + c_3, \cdots, a_{2k-3} + c_{2k-2}, a_{2k-2} + \gamma_3, a_{2k-1} + c_{2k-1}, \alpha_4 + \gamma_4),$$

$$\bar{\mathbf{a}} + \bar{\mathbf{c}} = (\bar{\alpha}_1 + \bar{\gamma}_1, \bar{\alpha}_2 + \bar{\gamma}_2, \bar{\alpha}_3 + \bar{\gamma}_3, \bar{\alpha}_4 + c_1, a_1 + c_2, \cdots, a_{2k-4} + c_{2k-3}, a_{2k-3} + c_{2k-2}, a_{2k-2} + \bar{\gamma}_4, a_{2k-1} + c_{2k-1}).$$

If $\mathbf{a} + \mathbf{c} \neq \bar{\mathbf{a}} + \bar{\mathbf{c}}$, by (17) and (19), we have $\ell(\mathbf{a} + \mathbf{c}, \bar{\mathbf{a}} + \bar{\mathbf{c}}) \geq 2k$, then conclusion follows. If $\mathbf{a} + \mathbf{c} = \bar{\mathbf{a}} + \bar{\mathbf{c}}$, we have $\mathbf{a} + \mathbf{c} = (0, 0, 0, \underbrace{1, 1, \cdots, 1}_{2k-2}, 0, 0)$ or $(1, 1, 1, \underbrace{0, 0, \cdots, 0}_{2k-2}, 1, 1)$, then $\ell(\mathbf{a} + \mathbf{b}, \mathbf{a} + \mathbf{c}) \geq 2k$ or $\ell(\mathbf{b} + \mathbf{c}, \mathbf{h}) \geq 2k$. So, we can always obtain two distinct codewords of $\mathcal{C}$ such that the length of their common subsequence is at least $2k$. $\qquad \square$

By Lemma 5.1, Lemma 5.2 and the proof of Corollary 3.2 we have the main theorem in this section.

Theorem 5.3 *Let $\mathcal{C}$ be an $[n, k]$ linear insdel code over $\mathbb{F}_2$.*

(1) *If $n > 2k$, $\mathbf{1} \notin \mathcal{C}$ and $\mathcal{C}$ is optimal with respect to the strict half-Singleton bound proposed in Theorem 3.1, then its code length n and dimension k satisfy $2k + 1 \leq n \leq 2k + 2$.*

(2) *If $n \geq 2k$ and $\mathcal{C}$ is optimal with respect to the half-Singleton bound, then its code length n and dimension k satisfy $2k \leq n \leq 2k + 1$.*

Let $\mathcal{C}$ be an $[n, k]$ linear insdel code. When $k = 2$, if $\mathbf{1} \notin \mathcal{C}$, there are 17 optimal $[5, 2]$ linear codes with respect to the strict half-Singleton bound given in Theorem 3.1; If $\mathbf{1} \in \mathcal{C}$ then there are 2 optimal $[4, 2]$ linear codes with respect to the half-Singleton bound. All these optimal linear insdel codes are listed in Table 1 and Table 2 by generators, respectively. A large number of experimental results show that Theorem 5.3 can be strengthened into the following conjecture.

Conjecture 5.4 *Let $\mathcal{C}$ be an $[n, k]$ linear insdel code over $\mathbb{F}_2$.*

(1) *If $n > 2k$, $\mathbf{1} \notin \mathcal{C}$ and $\mathcal{C}$ is optimal with respect to the strict half-Singleton bound in Theorem 3.1, then $\mathcal{C}$ has the parameters $[2k + 1, k, 4]$.*

(2) *If $n \geq 2k$ and $\mathcal{C}$ is optimal with respect to the half-Singleton bound, then $\mathcal{C}$ has the parameters $[2k, k, 4]$.*

Table 1: two generators of optimal [5,2] linear insdel codes

$\mathbf{v_1}$, $\mathbf{v_2}$	$\mathbf{v_1}$, $\mathbf{v_2}$	$\mathbf{v_1}$, $\mathbf{v_2}$
$(1,1,0,0,0),(0,0,1,1,0)$	$(1,1,0,0,0),(0,0,1,0,1)$	$(1,1,0,0,0),(0,0,0,1,1)$
$(1,0,1,0,0),(0,0,0,1,1)$	$(1,0,0,1,0),(0,1,1,0,0)$	$(1,0,0,0,1),(0,1,1,0,0)$
$(1,0,0,0,1),(0,1,0,1,0)$	$(1,0,0,0,1),(0,0,1,1,0)$	$(0,1,1,0,0),(0,0,0,1,1)$
$(0,1,0,0,1),(0,0,1,1,0)$	$(1,0,1,1,0),(0,0,1,1,1)$	$(1,1,0,0,1),(0,0,1,1,1)$
$(1,1,1,0,0),(0,0,1,1,1)$	$(1,1,0,1,0),(0,1,0,1,1)$	$(0,1,1,0,1),(1,0,0,1,1)$
$(1,1,1,0,0),(0,1,1,0,1)$	$(1,1,0,0,1),(1,0,1,1,0)$	

Table 2: two generators of optimal [4,2] linear insdel codes

$\mathbf{v_1}$, $\mathbf{v_2}$	$\mathbf{v_1}$, $\mathbf{v_2}$
$(1,1,0,0),\quad(0,0,1,1)$	$(1,0,0,1),\quad(0,1,1,0)$

6 Strict direct upper bound

In this section we prove the strict direct upper bound. This bound is only true for linear insdel codes. For a linear $[n, k]$ code $\mathcal{C} \subset \mathbb{F}_q^n$, the subset $S \subset \{1, \cdots, n\}$ of h coordinate positions is called an information free coordinate subset if the natural projection $\Phi_S : \mathcal{C} \longrightarrow \mathbb{F}_q^h$ defined by $\Phi_S((c_1, \cdots, c_n)) = (c_{i_1}, \cdots, c_{i_h})$ is surjective. It is clear $h \leq k$. When $h = k$ this is the information set. It is obvious that for any generator $k \times n$ matrix G of this linear $[n, k]$ code, the columns at these positions of an information-free subset are linear independent vectors in $\mathbb{F}_q^k$.

Theorem 6.1 (Strict direct upper bound) *Let $\mathcal{C} \subset \mathbb{F}_q^n$ be a linear $[n, k]$ code with the minimum Hamming distance d_H. Let $\mathbf{x} \in \mathcal{C}$ be a minimum weight codeword with its zero coordinate position*

set $[n] - \text{supp}(\mathbf{x}) = \{i_1, i_2, \cdots, i_{n-d_H}\}$, where $i_1 < i_2 < \cdots < i_{n-d_H}$. Suppose there are t pairs of coordinate positions $\{j_u, w_u\}$, $u = 1, \cdots, t$, satisfying $\{j_u, w_u\}$ is in some $[i_v + 1, i_{v+1} - 1]$ for $u = 1, \cdots, t$, and $\{j_1, w_1, \ldots, j_t, w_t\}$ is an information-free subset. Then

$$d_I(\mathcal{C}) \leq 2(d_H - t).$$

Proof. Let G be an $k \times n$ generator matrix of this linear code $\mathcal{C}$, with n columns $G_1, \ldots, G_n$. Let $\mathbf{x} = \mathbf{u} \cdot G$ be the minimum weight codeword claimed in the condition. Then $\mathbf{u}$ is a non-zero vector in $\mathbb{F}_q^k$ and $\mathbf{u} \cdot G_j = 0$ for $j = i_1, \ldots, i_{n-d_H}$. Since $\{j_1, w_1, \ldots, j_t, w_t\}$ is an information-free subset, $G_{j_1}, G_{w_1}, \ldots, G_{j_t}, G_{w_t}$ are linear independent vectors in $\mathbb{F}_q^k$. Then $G_{j_1} - G_{w_1}, G_{j_2} - G_{w_2}, \ldots, G_{j_t} - G_{w_t}$ are linear independent vectors in $\mathbb{F}_q^k$. Hence we can find a non-zero vector $\mathbf{v} \in \mathbb{F}_q^k$ satisfying $\mathbf{v} \cdot (G_{j_u} - G_{w_u}) = \mathbf{u} \cdot G_{w_u}$ for $1 \leq u \leq t$. Now for two codewords $\mathbf{x}_1 = \mathbf{v} \cdot G$ and $\mathbf{x}_2 = (\mathbf{v} + \mathbf{u}) \cdot G = \mathbf{x}_1 + \mathbf{x}$, their coordinates at positions $i_1 < i_2 < \cdots < i_{n-d_H}$ are the same. Since

$$\mathbf{v} \cdot G_{j_u} = (\mathbf{v} + \mathbf{u}) \cdot G_{w_u}, \quad u = 1, 2, \cdots, t,$$

the coordinates of $\mathbf{x}_1$ and $\mathbf{x}_2$ at position pairs $\{j_u, w_u\}$ are the same. Since $\{j_u, w_u\}$ is always in some $[i_v + 1, i_{v+1} - 1]$, we have a common subsequence with length $n - d_H + t$ of these two codewords. The conclusion is proved. $\square$

If a linear code $\mathcal{C}$ is projective then we have the following corollary.

Corollary 6.2 *Let $\mathcal{C} \subset \mathbb{F}_q^n$ be a projective linear code with the minimum Hamming distance $d_H > \frac{n+1}{2}$. Then*

$$d_I(\mathcal{C}) \leq 2(d_H - 1).$$

Proof. Since $n - d_H + 1 < d_H$, for any minimum weight codeword $\mathbf{x}$ with zero-coordinate positions $i_1 < i_2 < \cdots < i_{n-d_H}$, we have an interval $[i_v + 1, i_{v+1} - 1]$ containing at least two support coordinate positions of $\mathbf{x}$. The two columns at these two positions are linear independent from the condition $d_H(\mathcal{C}^\perp) \geq 3$. The conclusion follows directly. $\square$

Next we give two examples of linear insdel codes attaining the strict direct upper bound and an example showing that Theorem 6.1 is not true for nonlinear insdel codes.

Example 6.3 *Let $\mathcal{C}$ be an $[11, 4]$ linear code over $\mathbb{F}_2$ with the following generator matrix,*

$$G = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} = (G_1, G_2, \cdots, G_{11}).$$

One can verify that $d_H(\mathcal{C}) = 4$, and there is a minimum weight codeword $\mathbf{x} = (0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0)$ in $\mathcal{C}$ with a zero coordinate position set $\{1, 2, 3, 6, 7, 10, 11\}$. Then there are two pairs of coordinate positions $\{4, 5\}$ and $\{8, 9\}$ are in $[3, 6]$ and $[7, 10]$, respectively, such that $\{4, 5, 8, 9\}$ is an information-free subset. Then by Theorem 6.1, we know that $d_I(\mathcal{C}) \leq 2(d_H - 2) = 4$. On the other hand, we can verify that $d_I(\mathcal{C}) = 4$. In fact, there are two distinct codewords $\mathbf{x_1} = (0, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1)$ and $\mathbf{x_2} = \mathbf{x_1} + \mathbf{x} = (0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1)$ in $\mathcal{C}$ such that $\ell(\mathbf{x_1}, \mathbf{x_2}) = 9$, which is the longest common subsequence of codewords in $\mathcal{C}$. So, $d_I(\mathcal{C}) = 4$ and $\mathcal{C}$ attains the strict direct upper bound proposed in Theorem 6.1 for $t = 2$.

17

Example 6.4 *Let $\mathcal{C}'$ be a nonlinear code consisting of four codewords of $\mathcal{C}$ in Example 6.3 as follows:*

$$\mathcal{C}' = \big\{(0,0,0,0,0,0,0,0,0,0,0),\ (1,1,1,1,0,1,0,0,0,1,1),$$
$$(1,0,0,0,0,0,0,0,1,1,1),\ (0,0,0,1,1,0,0,1,1,0,0)\big\}.$$

One can verify that $d_H(\mathcal{C}') = 4$ and $d_I(\mathcal{C}') = 8$. A minimum weight codeword $(1,0,0,0,0,0,0,0,1,1,1)$ in $\mathcal{C}'$ has a zero coordinate position set $\{2,3,4,5,6,7,8\}$. It is easy to see that there is a pair of coordinate positions $\{9,11\}$ in $[9,11]$ such that this set is an information-free subset. However, $d_I(\mathcal{C}') > 2(d_H(\mathcal{C}') - 1)$. This example shows that Theorem 6.1 is not true for nonlinear insdel codes.

Example 6.5 *Let p be a prime number and let $e > 1$ be a positive integer. Let $i_j = 2^{j-1}$ for $1 \leq j \leq n$ satisfying $3 \cdot 2^{n-2} < e$. Let θ be a primitive element in the finite field $\mathbb{F}_{p^e}$ and*

$$\mathcal{C} = \big\{(\lambda + \mu\theta^{i_1}, \lambda + \mu\theta^{i_2}, \cdots, \lambda + \mu\theta^{i_n}) \,|\, \lambda, \mu \in \mathbb{F}_{p^e}\big\}$$

be a two-dimensional RS code of length n over $\mathbb{F}_{p^e}$. Since $\mathcal{C}$ is an MDS code, $d_H(\mathcal{C}) = n - 1$. From Corollary C in [4], we know that $d_I(\mathcal{C}) = 2n - 4$. Thus, $d_I(\mathcal{C}) = 2(d_H - 1)$ and $\mathcal{C}$ attains the strict direct upper bound proposed in Corollary 6.2.

7   Concluding remark

In this paper, we proposed the strict half-Singleton bound for linear insdel codes without all 1 codeword and a method to construct optimal linear insdel codes with respect to this upper bound. Then, we proved that the length of optimal binary linear insdel codes with respect to the (strict) half-Singleton bound is about twice the dimension. A large number of experimental results suggested that optimal binary linear insdel codes have parameters $[2k, k, 4]$ or $[2k + 1, k, 4]$ with respect to the half-Singleton bound or the strict half-Singleton bound proposed in Theorem 3.1, respectively. Moreover, interestingly explicit optimal linear insdel codes attaining the (strict) half-Singleton bound, with the code length being independent of the finite field size, were obtained. Finally, we also gave the strict direct upper bound for the minimum insdel distances of linear insdel codes and optimal linear insdel codes attaining our strict direct upper bound were presented.

References

[1] K. A. S. Abdel-Ghaffar, H. C. Ferreira, L. Cheng, Correcting deletions using linear and cyclic codes. IEEE Trans. Inf. Theory, 56(10): 5223-5234, 2010.

[2] E. Brill, R.C. Moore, An improved error model for noisy channel spelling corrections, Proc. of the Thirty Eight Annual Meeting on Association for Computational Linguistics (ACL), 286-293, 2000.

[3] Y. Chee, H. Kiah, A. Vardy, V. Vu, E. Yaakobi, Codes correcting position errorsin racetrack memories, 2017 IEEE Information Theory Workshop (ITW), Kaohsiung, 161-165, 2017.

[4] B. Chen, G. Zhang, Improved Singleton bound on insertion-deletion codes and optimal constructions, IEEE Trans. Inf. Theory, 68(5): 3028-3033, 2022.

[5] H. Chen, Coordinate-ordering-free upper bounds for linear insertion-deletion codes, arXiv:2106.10782 [cs.IT], 2021, online version, IEEE Trans. Inf. Theory, 2022.

[6] K. Cheng, Z. Jin, X. Li, K. Wu, Deterministic document exchange protocols, and almost optimal binary codes for edit errors. IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS), 200-211, 2018.

[7] K. Cheng, V. Guruswami, B. Haeupler, X. Li, Efficient linear and affine codes for correcting insertions/deletions, Proc. 2021 ACM-SIAM Symposium on Discrete 33 Algorithms, SODA 2021, 1-20, SIAM, 2021.

[8] V. Chvátal, D. Sankoff, Longest common subsequences of two random sequences, J. Appl. Probability, 12: 306- 315, 1975.

[9] R. Con, A. Shpilka, I. Tamo, Linear and Reed-Solomon codes aganst adversarial insertions and deletions, arXiv:2107.05699v2 [cs.IT], 2021.

[10] T. Do Duc, S. Liu, I. Tjuawinata, C. Xing, Explicit constructions of two-dimensional Reed-Solomon codes in high insertion and deletion noise regime, IEEE Trans. Inf. Theory, 67 (5): 2808-2820, 2021.

[11] R. Gabrys, F. Sala, Codes correcting two deletions, IEEE Trans. Inf. Theory, 65(2): 965-974, 2019.

[12] S.W. Golomb, J. Dsvey, I. Reed, H. Van Trees, J. Stiffler, Synchronization, IEEE Transactions on Communications Systems, 11(4): 481-491, 1963.

[13] B. Haeupler, A. Shahrasbi, Synchronization strings: codes for insertions and deletions approaching the Singleton bound, Proc. of the 49th Annual ACM Symposium on Theory of Computing (STOC), 33-46, 2017.

[14] B. Haeupler, A. Shahrasbi, M. Sudan, Synchronization Strings: List Decoding for Insertions and Deletions, 45th International Colloquium on Automata, Languages and Programming(ICALP), 2018.

[15] B. Haeupler, Optimal document exchange and new codes for insertions and deletions, IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS), 334-347, 2019.

[16] B. Haeupler, A. Shahrasbi, Synchronization strings and codes for insertions and deletions: A Survey, IEEE Trans. Inf. Theory, 67(6): 3190-3206, 2021.

[17] S. Jain, F. Hassanzadeh, M. Schwartz, J. Bruck, Duplication-correcting codes for data storage in the DNA of living organisms, IEEE Trans. Inf. Theory, 63(8): 4996-5010, 2017.

[18] V. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals, Doklady Akademii Nauk SSSR, 163: 845-848, 1965.

[19] A. Lenz, P. Siegal, A. Wachter-Zeh, E. Yaakobi, Codes over sets for DNA storage, IEEE Trans. Inf. Theory, 66(4): 2331-2351, 2020.

[20] L. McAven, R. Safavi-Naini, Classification of the deletion correcting capabilites of Reed-Solomon codes of dimension 2 over prime fields, IEEE Trans. Inf. Theory, 53(6): 2280-2294, 2007.

[21] F.J. Och, Minimum error rate training in statistical machine translation, In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics(ACL), Association for Computational Linguistics, Stroudsburg, PA, USA., 160-167, 2003.

[22] D. Sankoff, J.B. Kruskal, editors, Time warps, string edits, and macromolecules: the theory and practice of sequence comparison, Addison-Wesley Pub. Comp., Advanced Book Program, 1983.

[23] C. Schoeny, A. Wachter-Zeh, R. Gabrys, E. Yaakobi, Codes correcting a burst of deletions or insertions, IEEE Trans. Inf. Theory, 63(4): 1971-1986, 2017.

[24] E. Tanaka, T. Kasai, Synchronization and substitution error-correcting codes for the Levenshtein metric, IEEE Trans. Inf. Theory, 22(2): 156-162, 1976.

[25] D. Tonien, R. Safavi-Naini, Construction of deletion correcting codes using generalized Reed-Solomon codes and their subcodes. Des. Codes Cryptogr., 42(2): 227-237, 2007.

[26] Y. Wang, L. McAven, R. Safavi-Naini, Deletion correcting using generalized Reed-Solomon codes, Progress in Computer Science and Applied Logic, 23: 345-358, 2004.

[27] R. Xu, D. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw., 16(3): 645-678, 2005.